

Sentiment Analysis with Orange

Amy Larner Giroux, PhD

UCF Center for Humanities & Digital Research

NEH Digital Culture Summer Institute



Welcome

The tutorial outlined in this slide deck is an extension of the Introduction to Data Analysis in Orange. We will work through your second workflow in Orange and learn about Sentiment Analysis. For full details, please consult the “Orange Tweet Analysis Tutorial” PDF available in Google Classroom. That document covers both tutorials for this week in more detail.

Agenda

- Downloading the dataset
- Beginning where we left off in the previous tutorial
- Detailed steps for your second workflow – from Preprocess Text to Sentiment Analysis
- Updating the current widgets with new data
- Adding a Twitter Profiler and Box Plot
- Adding Sentiment Analysis and Scatter Plots
- Your NEH Institute deliverables for this tutorial



Tutorial Dataset

Please download the dataset:

<http://chdr.cah.ucf.edu/neh-digculture/Tweet-Profiled-ReadyForOrangeNEH.xlsx>

The dataset for this tutorial has 4 categorical columns in addition to the regular tweet data.

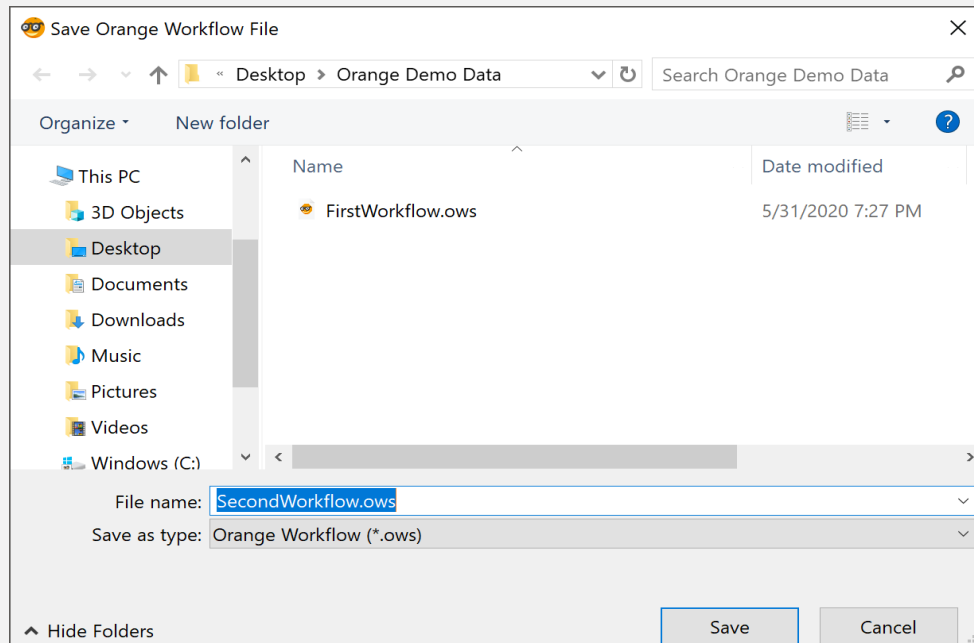
The criteria decisions for the Influentials, Opinion-leaders, and Political Leaning columns is discussed in the “Orange Tweet Analysis Tutorial” PDF. The last column, Political Leaning (numeric), was added to allow access to the category within the scatter plot widget. More on that later in the tutorial.

	A	B	C	D	E	F	G	H	I	J	K	L
	Date	Text	User	To	Retweets	Favorites	Mentions	Hashtags	Influentials	Opinion-leaders	Political Leaning	Political Leaning (numeric)
1												
2	3/23/2020	.@realDonaldTrump just said: "if it were up to the doctors, they'd shut down the entire world."	DrLeanaWen		24641	100067	@realDonaldTrump	#covid19	yes	yes	left	1
3	3/16/2020	There it is. I've been deathly afraid of this exact moment where Trump turns to racism and	eugenegu	realDonaldTrump	24532	216814			yes	no	left	1
4	3/19/2020	America's Private Sector is stepping up to help us be STRONG! Many of the Nation's distillers, large	realDonaldTrump		20931	86850		#COVID19	yes	yes	right	-1
5	3/12/2020	WOW. Rep. Porter just read the testing costs to CDC Director Redfield, who did not know them, and	girlsreallyrule		18318	57843	@realDonaldTrump	#Covid_19	yes	yes	left	1
6	3/21/2020	For anyone who read the tweet from @realDonaldTrump saying hydroxychloroquine & azithromycin	tedlieu	EdselSal	17937	37171	@realDonaldTrump	#Covid_19	yes	yes	left	1
7	3/15/2020	Thank you @realDonaldTrump for your leadership. The task force is tackling real national security	SaraCarterDC	realDonaldTrump	16636	68270	@realDonaldTrump		yes	yes	right	-1
8	3/14/2020	Attending meetings on covid19 in the White House. Working with States and local governments,	realDonaldTrump		16218	98222			yes	yes	right	-1
9	3/14/2020	Just had a nice conversation with Prime Minister @JustinTrudeau of Canada. Great to hear that his	realDonaldTrump		15174	93994	@JustinTrudeau		yes	yes	right	-1

Reopen Your First Workflow

Open Orange and use **Ctrl-Alt-O** or **File -> Open and Freeze** to open your workflow from the previous tutorial.

Using the previous workflow will save us some setup time.



Once the workflow is open:

Click on File -> Save As and name it **SecondWorkflow.ows** as we do not want to overwrite your first project.



File Load

Double-click on the **File** widget to open it.

Using the **open folder button**, browse and choose the tutorial spreadsheet. If you are on Windows and have saved it to your desktop, you may need to navigate through Users -> your user name -> Desktop to get to the file.



After opening the tutorial spreadsheet, click on the **Reload** button at the top right to refresh the data.

You can close this window after the data loads (matching this screen shot.)

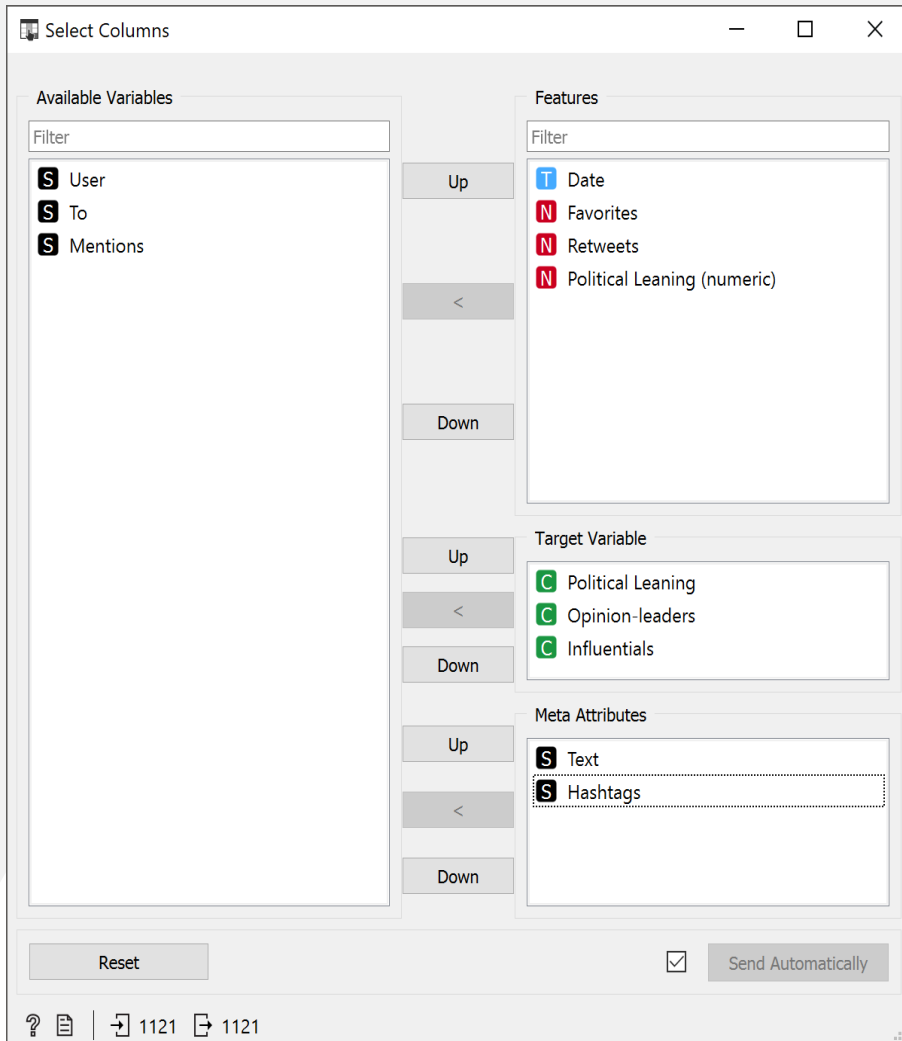
The screenshot shows the 'File' widget interface in Orange Data Mining. The 'File' tab is selected, showing the loaded file 'Tweet-Profiled-ReadyForOrangeNEH.xlsx'. The 'Info' section indicates 1121 instances, 7 features, and 5 meta-attributes. The 'Columns' section displays a table of the loaded data features.

	Name	Type	Role	Values
1	Date	datetime	feature	
2	Retweets	numeric	feature	
3	Favorites	numeric	feature	
4	Influentials	categorical	feature	no, yes
5	Opinion-leaders	categorical	feature	no, yes
6	Political Leaning	categorical	feature	left, neutral, right
7	Political Leanin...	numeric	feature	
8	Text	text	meta	

At the bottom, there are buttons for 'Browse documentation datasets', 'Reset', and 'Apply'. The status bar at the very bottom shows a question mark, document icon, and the instance count '1121'.

Select Columns

Double-click on the **Select Columns** widget to open its options screen.



You may **drag and drop** the fields into the different areas of the screen or highlight them and use the arrows in the center section.

Move the categorical fields (those with the green “C” icon) to the Target Variable area.

The remainder should be in the places they were from the first tutorial.

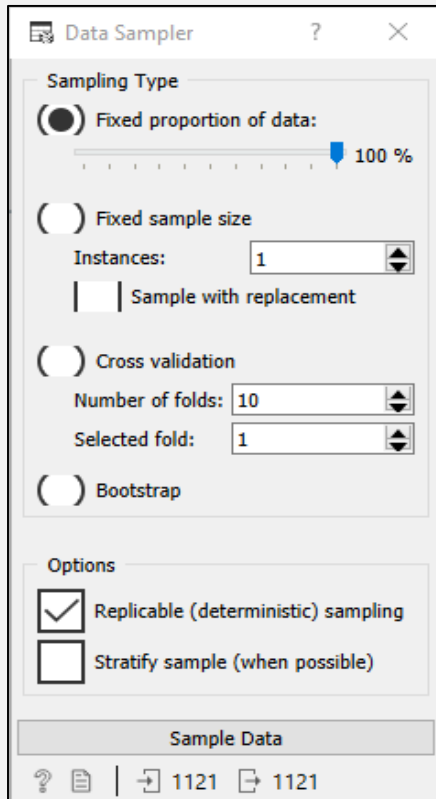
Once the column names have been arranged in the correct areas, you can close this screen.



Data Sampler

Double-click the Data Sampler widget.

For this tutorial, we are going to use the entire dataset. We could just remove this widget and connect Select Columns to Corpus, but you may want to use this workflow for different larger files in the future.



Use the **Fixed proportion of data** slider and set it to 100%.

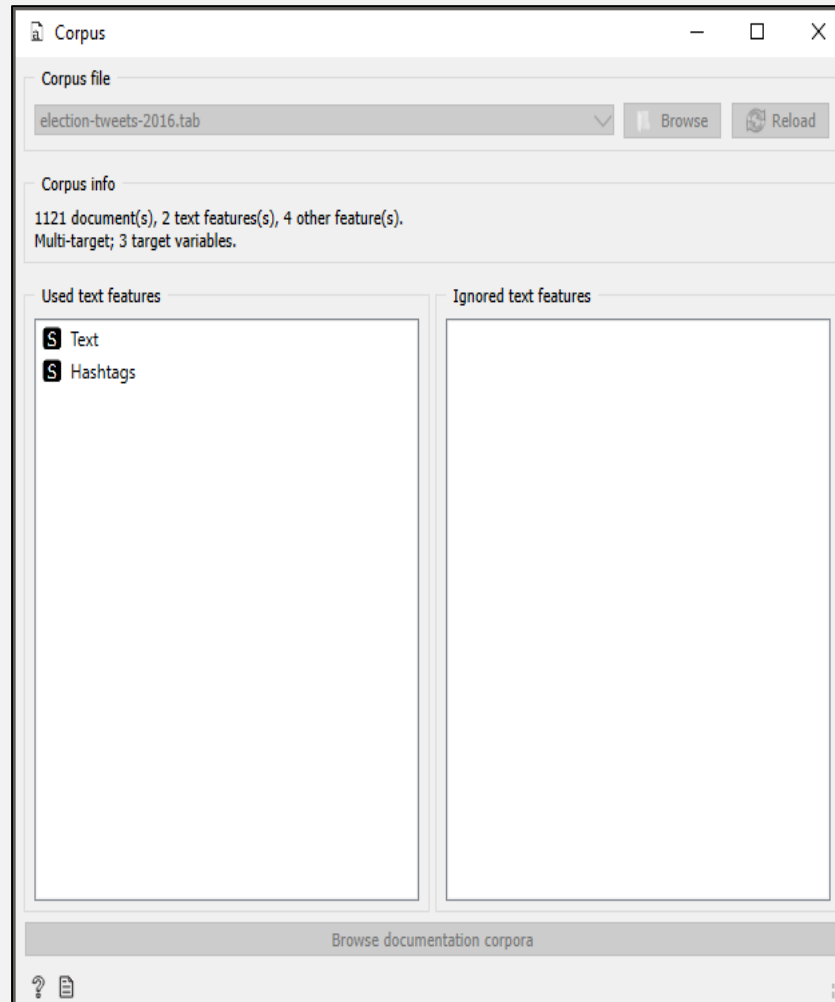
Remember to **click** on the **Sample Data** button to resample at the full amount.

Once you have sampled your data, you can close this screen



Corpus

Double-click on the Corpus widget to open the options screen.



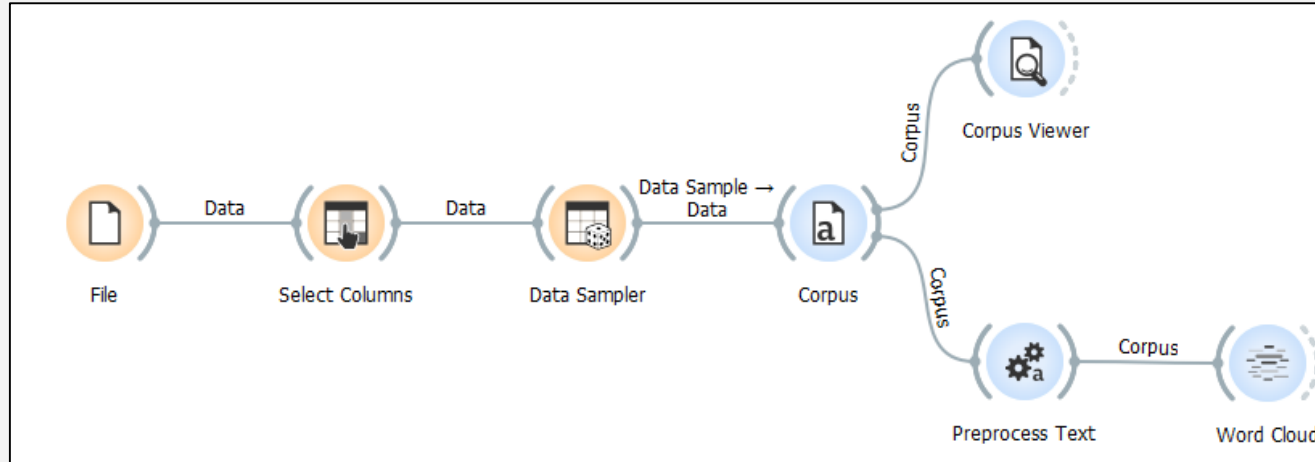
Both fields should be in the
Used text features column.

Once both fields are in the left
column, close this window.



Workflow Check

At this point in the tutorial, your workflow should resemble the one in this image; the same as the end of the last tutorial.



We will leave the Preprocess Text widget alone unless you changed the Stopwords language to something other than English. If so, please set it back to English.

We have imported the new COVID19 data file; chosen the columns for target variables; changed the Data Sample widget to use 100% of our dataset; and made sure the Corpus had the columns correctly set.

We are now ready for sentiment analysis.

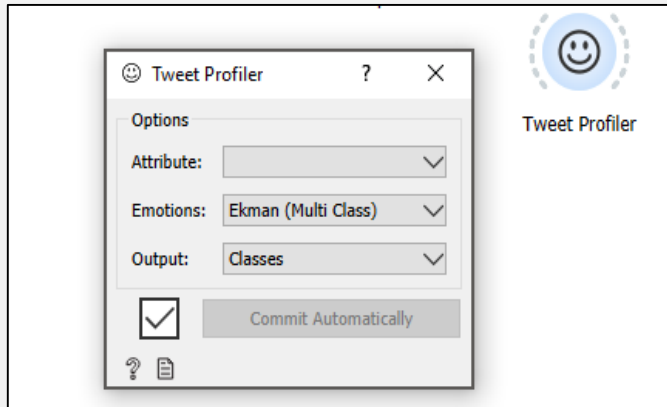


Tweet Profiler

Our first look at a sentiment analysis will be with the Tweet Profiler widget.

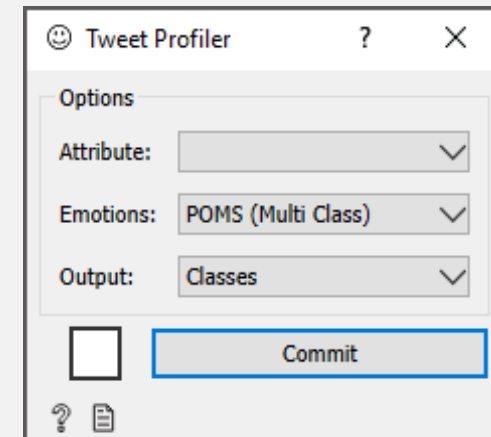
Add it to your work area, but **DO NOT** connect it to any other widget yet. This is the kind of widget that processes immediately and we want to change some settings first.

Double-click the Tweet Profiler to open the options screen.



The profiler defaults to the Ekman (Multi Class) model for emotions and is set to Commit Automatically.

In this tutorial we will be using the **POMS (Multi Class)**¹, so please **select** that and **uncheck** the box so that the button now says “Commit.”



¹ The assigned reading for the homework, “Emotion Recognition on Twitter: Comparative Study and Training a Unison Model” by Niko Colneriĉ and Janez Demsar describes in detail the various models used in the Tweet Profiler widget.